ENT-ICIPATE

Checkpoint #3

Team 7: Alessandro Carrabs, Xiao Quan Ji, Samesun Singh

FONDAZIONE links

1859

SALUS SCIENTIA

AZIENDA OSPEDALIERO-UNIVERSITARIA
Città della Salute e della Scienza di Torino

TABLE OF CONTENTS

OVERVIEW

OBJECTIVES

WE WANT TO:

- BUILD A ML MODEL TO ESTIMATE EACH PATIENT'S RISK OF SOME COMPLICATIONS (NOSOCOMIAL INFECTION AND PHARYNGO-/ORO-CUTANUOUS FISTULA)

- COMPARE DIFFERENT MODELS AND STRATEGY TO IDENTIFY A ROBUST AND INTERPRETABLE SOLUTION

UNNECESSARY ALERTS ARE UNPLEASANT.

STILL, IT'S BETTER THAN MISSING A TRUE COMPLICATION.

Dr. Emma Collins

ENT Surgeon

VALUE PROPOSITION

Transforms raw clinical data from 550+ ENT oncology patients into actionable insights, enabling earlier identification of high-risk cases and improving post-surgical safety.

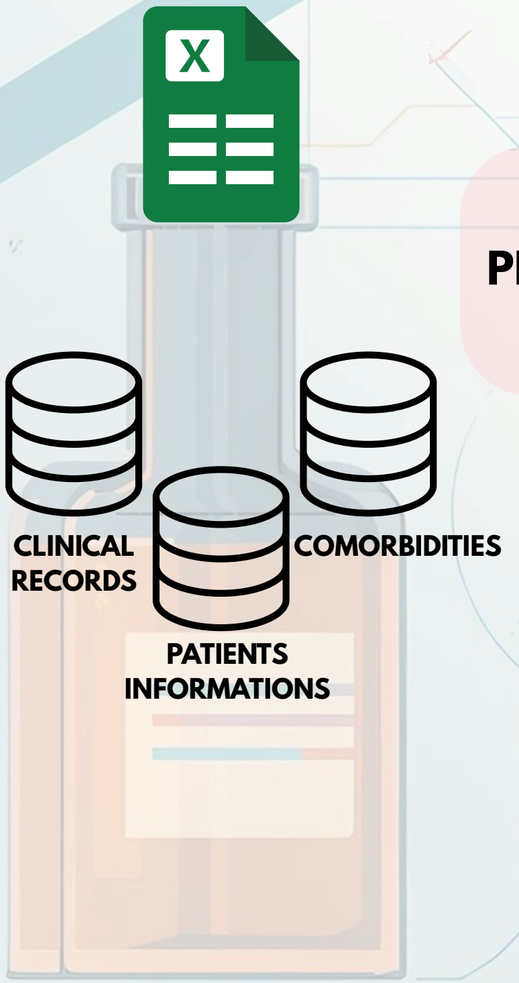
Builds a data-driven infrastructure that standardizes heterogeneous medical records, integrates them into a predictive pipeline, and lays the foundation for scalable AI-assisted clinical workflows.

RESEARCH QUESTIONS

CAN MACHINE LEARNING MODELS RELIABLY PREDICT POST-SURGICAL COMPLICATIONS IN ENT ONCOLOGY, DESPITE THE RARITY OF THESE EVENTS?

WHAT ARE THE MAIN CLINICAL AND SURGICAL PREDICTORS OF POST-OPERATIVE COMPLICATIONS?

FUNCTIONAL DIAGRAM



PREPROCESSING

FEATURE SELECTION

MACHINE LEARNING MODELS

RANDOM FOREST

LOGISTIC REGRESSION

GRADIENT BOOSTING

CATBOOST

OUTPUT

POST SURGERY COMPLICATIONS

PREDICT POSSIBLE NOSOCOMIAL INFECTION

EXPLAINABLE OUTPUT (FOR ENT-DOCTORS

IDENTIFY FEATURES THAT INCREASE OR DECREASE PREDICTED RISK, EVEN FOR RARE COMPLICATIONS

DATASET AND DATA PREPROCESSING

COMPOSITION OF THE DATASET

574 Patients

64 Features

- **Patients demographics and habits**
- **Comorbidities**
- **Surgical and operative treatments**
- **Targets (binary)**
  - **Fistula**
  - **Nosocomial infection**

Clinical Data Challenges

- High heterogeneity across variables
- Many categorical features stored as free text
- TNM staging unstructured and inconsistently encoded
- Hidden missing values not immediately detectable

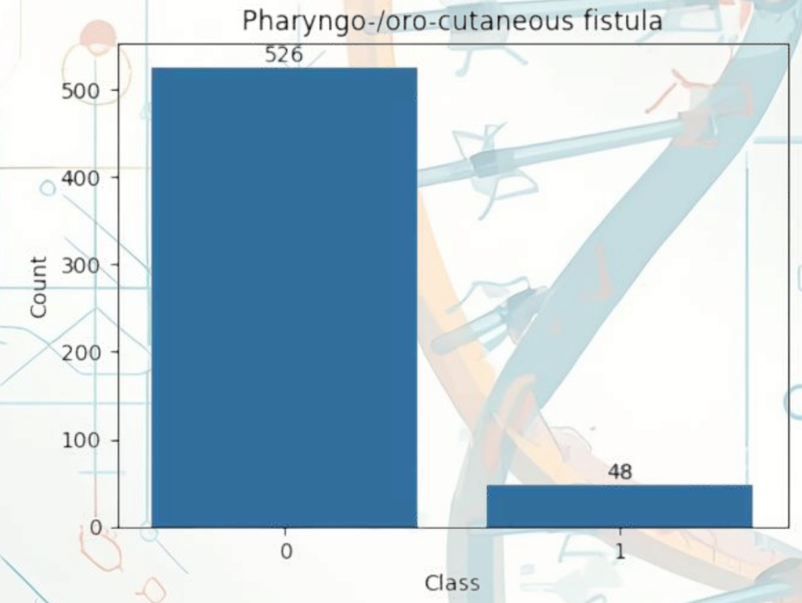
IMBALANCE ASSESTMENT

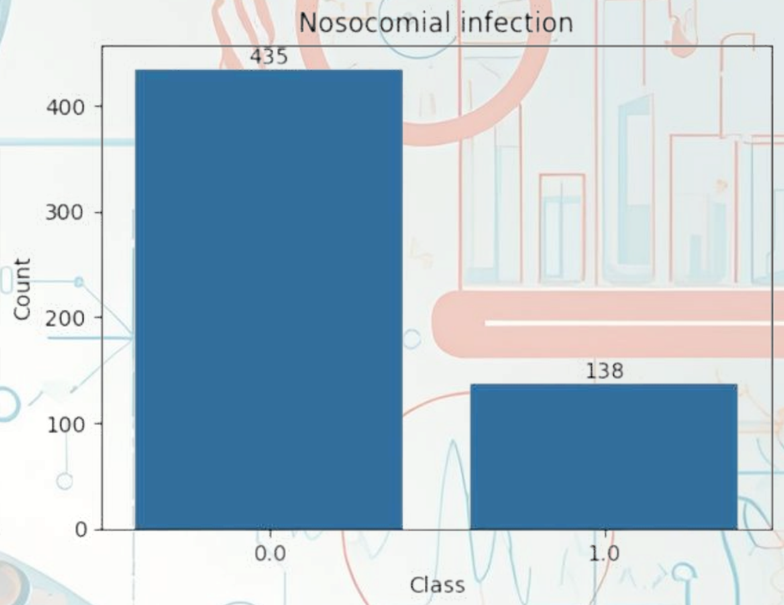
Plots reflect the rarity of post-operative complications in real clinical scenario (0 = no complication, 1 = complication)

This creates a majority-class bias, models may appear accurate by predicting mostly 0s, but fail to detect the true positives.

To address this we:

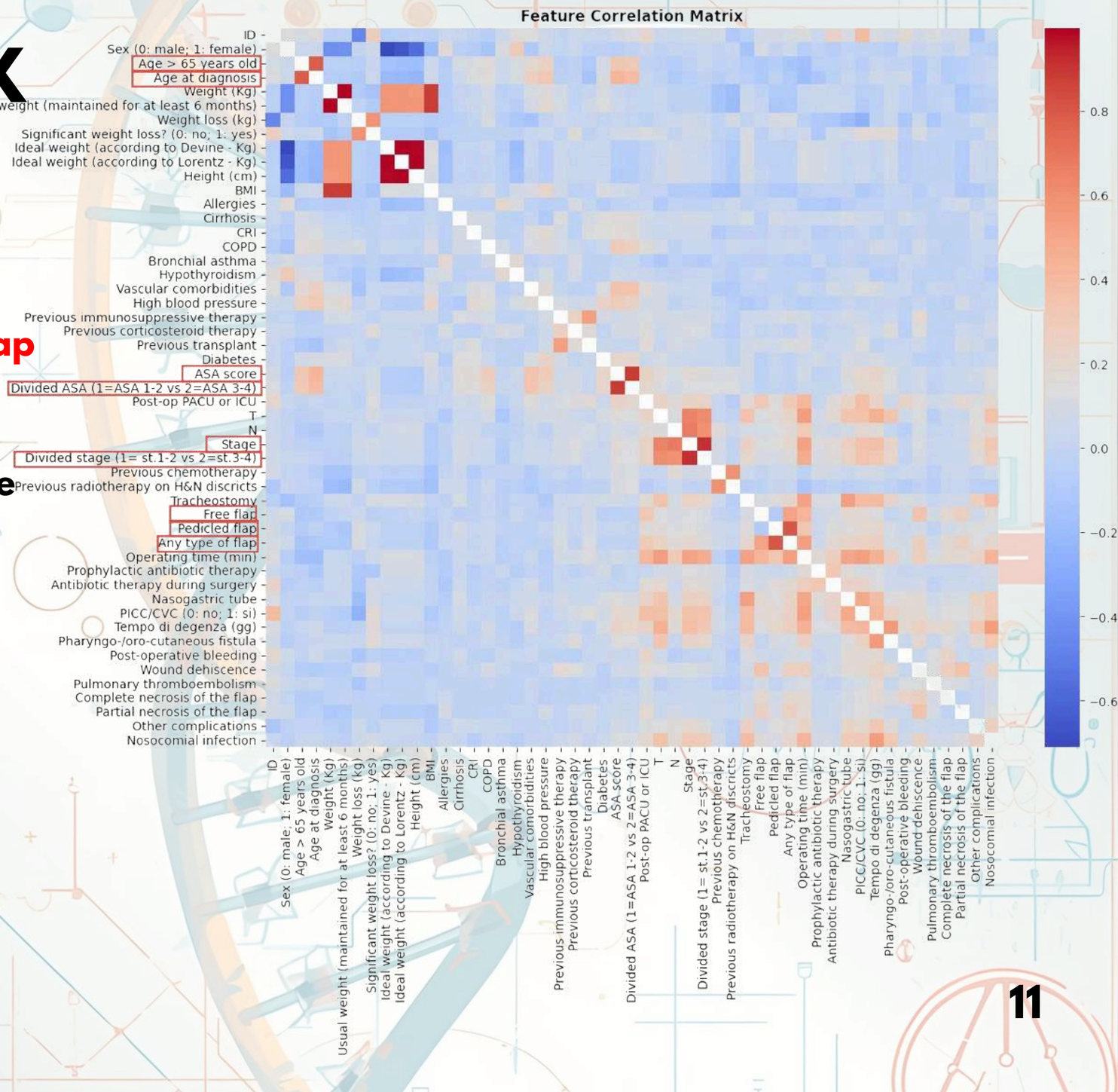
- **evaluate using class-1 metrics (Recall/Precision/F1)**
- **tune decision thresholds on validation to prioritize sensitivity (safety-first).**





Goal: maximize detection of high-risk patients (minimize false negatives), accepting some additional false alarms.

CORRELATION MATRIX



Some features capture the same clinical concept in different ways, creating overlap and redundancy.

Redundant features do not add predictive value and can reduce model stability.

EXAMPLES:

- **Weight-related information (height, weight, ideal weight)**
- **ASA and Stage scores encoded in multiple forms**
- **Age**
- **Presence of flap —> Keep only one version of each correlated feature to avoid multicollinearity**

OTHER PREPROCESSING STEPS

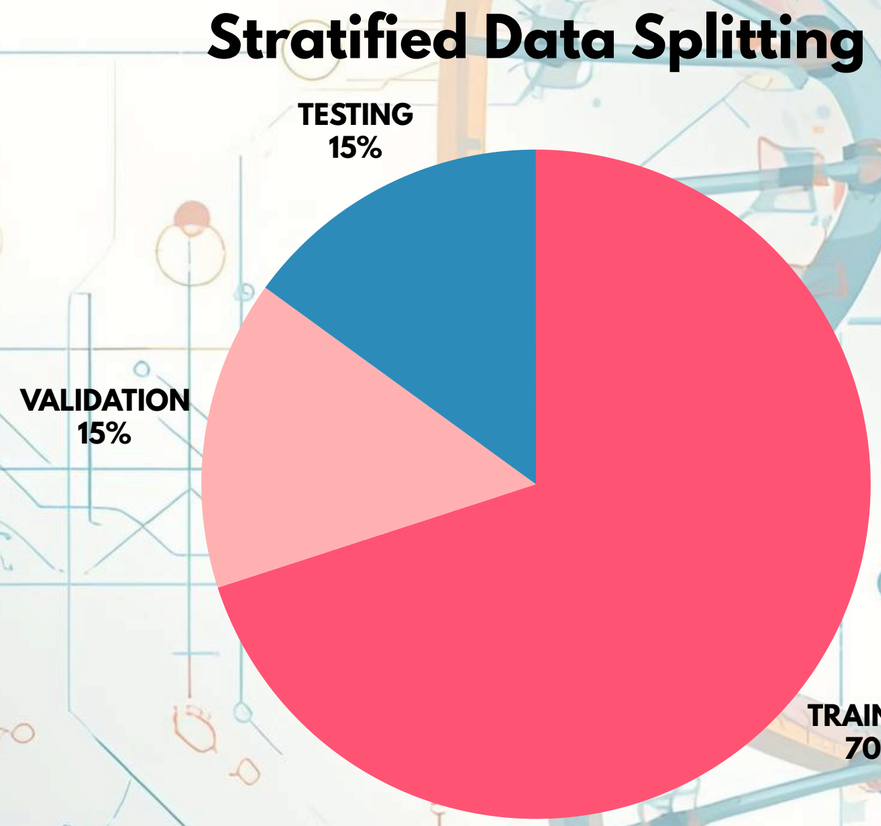
-**Standardized target labels**: Fixed inconsistent entries and converted text to numbers (e.g., “fistula…” → 1)

-**Recoded TNM staging**: Unified different formats (IIIA, 3a, IVB…) into 0–4 integers for consistency.

-**Handled missing values**: Removed rows missing critical clinical information (Stage, Operating time, Free flap, Antibiotic therapy…).
For clinically important features (e.g., significant weight loss), missing values were imputed with -1 to keep the information without introducing false assumptions

IMPLEMENTATION

DATA SPLIT



STRATIFIED: KEEPS THE **SAME CLASS DISTRIBUTION** IN BOTH TRAIN, VAL AND TEST SETS

MODELING PIPELINE

HYPERPARAMETER TUNING

5-fold stratified CV on training set
GridSearchCV optimized for PR-AUC (handles class imbalance)

MODEL SELECTION

Compared Logistic Regression, Random Forest, CatBoost, XGBoost on validation set.
Winner selected based on PR-AUC

THRESHOLD OPTIMIZATION

Tested multiple thresholds on validation set
Selected threshold maximizing F1-score for positive class

FINAL REFIT & EVALUATION

Refit winning model on train + validation sets
Evaluated on held-out test set using optimized threshold

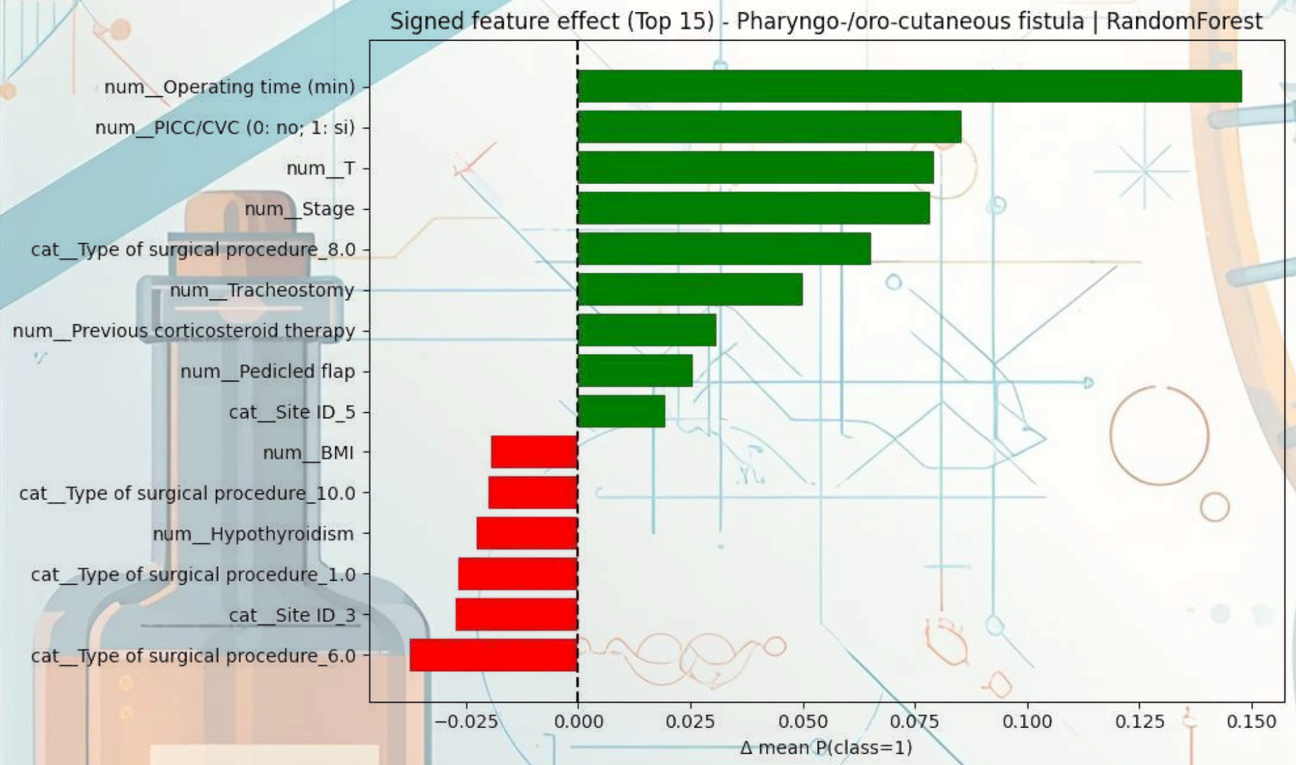
EVALUATION

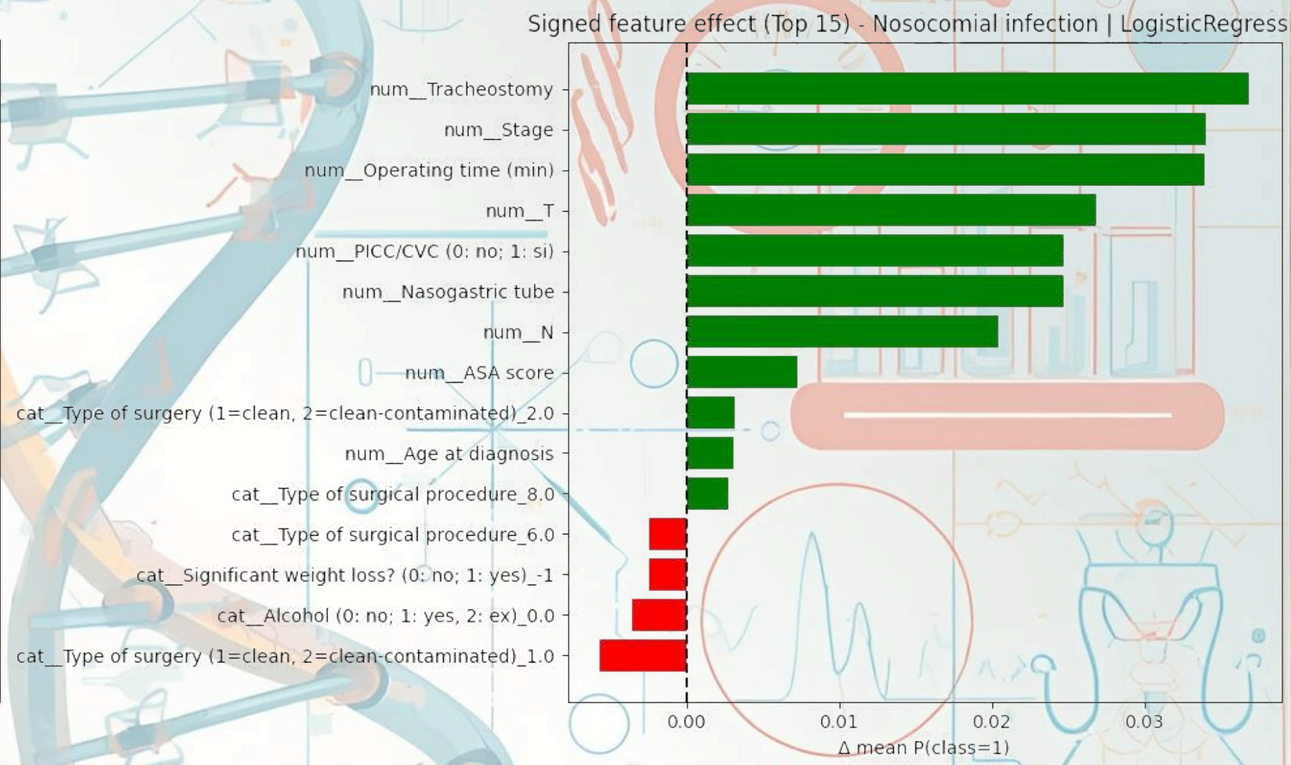
RESULTS ON TEST SET

EVALUATION METRIC	RANDOM FOREST for TARGET: Pharyngo-/oro-cutanoeous fistula	LOGISTIC REGRESSION for TARGET: Nosocomial infection
ROC-AUC	0.945	0.833
F1	0.571	0.604
Recall	0.857	0.762
Precision	0.429	0.5

For class “1”

SIGNED FEATURE EFFECT





For each feature, we measure the change in average predicted risk when moving from low to high

- **Positive effect (green) : increases predicted complication risk.**
- **Negative effect (red) : decreases predicted complication risk**

CONCLUSIONS

EXPLAINED RESULTS FOR STAKEHOLDERS

- Early risk stratification: identifies patients at higher risk of fistula & nosocomial infection
- Safety-first threshold: prioritizes high recall to minimize missed complications
- Actionable output: flagged patients can receive closer monitoring, preventive measures, or adjusted surgical plans.

- Standardized data pipeline: cleans heterogeneous clinical variables and ensures reproducible training & evaluation
- Resource allocation support: guides follow-up intensity, nursing workload, and post-op surveillance planning
- Scalable workflow: can be extended to new targets and future patient cohorts

LIMITATIONS AND FURTHER POSSIBLE IMPROVEMENTS

LIMITATIONS:

Rare events → metrics can be sensitive to a small number of cases (especially on the positive class).

False positives at safety-first thresholds → increased monitoring workload must be considered.

Single center dataset & heterogeneous encoding → external validation needed to confirm generalization.

NEXT STEPS:

External validation on **future patients** (robustness and generalizability).

Probability **calibration** → interpretable "risk scores" instead of raw probabilities.

Clinical operating **policy**: choose thresholds based on hospital capacity (safety-first vs workload).

Per-patient explainability: provide the top contributing factors for each flagged case to improve adoption.

THANK YOU

FOR THE ATTENTION

QUESTIONS?